Optimistic Value Instructors for Cooperative Multi-Agent Reinforcement Learning

Chao Li[1], Yupeng Zhang[2], Jianqi Wang[3], Yujing Hu[4]*, Shaokang Dong[1], Wenbin Li[1], Tangjie Lv[4], Changjie Fan[4], Yang Gao[1]*

[1] State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China
[2] Alibaba DAMO Academy, Hangzhou, China
[3] Meituan, Beijing, China
[4] NetEase Fuxi AI Lab, Hangzhou, China
chaoli1996@smail.nju.edu.cn, huyujing@corp.netease.com, gaoy@nju.edu.cn

Abstract

In cooperative multi-agent reinforcement learning, decentralized agents hold the promise of overcoming the combinatorial explosion of joint action space and enabling greater scalability. However, they are susceptible to a game-theoretic pathology called *relative overgeneralization* that shadows the optimal joint action. Although recent value-decomposition algorithms guide decentralized agents by learning a factored global action value function, the representational limitation and the inaccurate sampling of optimal joint actions during the learning process make this problem still. To address this limitation, this paper proposes a novel algorithm called *Optimistic Value Instructors (OVI)*. The main idea behind OVI is to introduce multiple optimistic instructors into the value-decomposition paradigm, which are capable of suggesting potentially optimal joint actions and rectifying the factored global action value function to recover these optimal actions. Specifically, the instructors maintain optimistic value estimations of per-agent local actions and thus eliminate the negative effects caused by other agents' exploratory or sub-optimal non-cooperation, enabling accurate identification and suggestion of optimal joint actions. Based on the instructors' suggestions, the paper further presents two instructive constraints to rectify the factored global action value function to recover these optimal joint actions, thus overcoming the RO problem. Experimental evaluation of OVI on various cooperative multi-agent tasks demonstrates its superior performance against multiple baselines, highlighting its effectiveness.

Introduction

Cooperative multi-agent reinforcement learning (MARL) has made significant progress in various real-world multi-agent tasks, including traffic signal control (Wang et al. 2020b), autonomous vehicles (Cao et al. 2012), sensor networks (Zhang and Lesser 2011), and robot swarm (Huang et al. 2020). However, learning effective coordinated policies for agents in these complex multi-agent tasks remains a major challenge. The scalability problem is a significant hurdle as the joint action space grows exponentially with the number of agents. Consequently, a centralized agent fails to address the combinatorial explosion of all agents' joint action space, while a decentralized agent conditioned on each agent's local information demonstrates greater scalability in practice (Claus and Boutilier 1998).

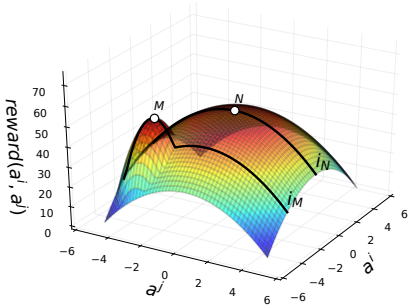

Figure 1: The relative overgeneralization pathology in a two-agent cooperative stage game.

Although decentralized agents allow for scalable learning of the joint coordinated policy by combining all agents' decentralized policies, they are susceptible to a game-theoretic pathology known as *relative overgeneralization (RO)*. This issue hinders agents from selecting the optimal joint actions. RO occurs when a sub-optimal Nash Equilibrium in the joint action space is preferred over an optimal Nash Equilibrium, which is because each agent's local action in the sub-optimal equilibrium is a better choice when matched with arbitrary actions of other collaborative agents (Wei and Luke 2016). For example, we consider a two-agent cooperative stage game. As depicted in Figure 1, the axes a^i and a^j respectively represent actions that agents i and j may select, and the axis $reward\ (a^i, a^j)$ denotes the joint reward after selecting the joint action (a^i, a^j). One can observe that the joint action M has higher reward than N. However, when agent j selects its action uniformly, the sub-optimal local action i_N has better average local value estimation than the optimal ones i_M and thus agent i tends to select action i_N, leading to a sub-optimal joint action.

The RO problem primarily stems from the limited information available to decentralized agents, as their decentralized value estimations of local coordinated actions may be biased by other agents' exploratory or sub-optimal non-cooperative action selections. One promising approach to al-

*Corresponding authors: Yujing Hu, Yang Gao

leviate it is to incorporate a centralized controller to guide the policy learning of multiple decentralized agents, referred to as the centralized training with decentralized execution (CTDE) paradigm. In this framework, value-decomposition algorithms (Sunehag et al. 2017; Rashid et al. 2020b; Son et al. 2019; Wang et al. 2020a) directly approximate the true joint action value function using factored versions that comprise all agents' decentralized utility functions. In addition, to ensure consistency between optimal local actions and the optimal joint actions derived from the factored global action value function, the Individual-Global-Max (IGM) principle (Son et al. 2019) is introduced. The underlying intuition is that if the true joint action value function is approximated accurately by the factored versions, the optimal joint actions can be identified by the factored ones and derived by local greedy operations on all agents' utility functions. However, value-decomposition algorithms often struggle to accurately identify the optimal joint actions and still suffer from the RO problem, which arises from either representational limitations or inaccurate sampling of optimal joint actions during the learning process.

To overcome above limitations, in this paper, we propose a novel algorithm called *Optimistic Value Instructors (OVI)*. The algorithm introduces multiple optimistic instructors to suggest potentially optimal joint actions and further presents two types of instructive constraints to rectify the learned factored global action value function. Specifically, we implement a per-agent optimistic instructor by learning an optimistic instructive value function for each agent. This value function effectively eliminates the negative effects caused by other agents' exploratory or sub-optimal non-cooperative action selections. Consequently, it enables the instructors to identify and suggest local coordinated actions for agents. Furthermore, based on the suggestions provided by instructors, we emphasize the importance of optimal joint actions when updating the factored global action value function for algorithms with representational limitations. For algorithms that achieve complete IGM expressiveness, we encourage agents to imitate their counterpart instructors. This imitation accelerates the learning of optimal joint policy.

We incorporate our algorithm into two state-of-the-art value-decomposition algorithms, QMIX and QPLEX. We then evaluate the performance of our algorithm by comparing it against multiple baselines in a variety of cooperative multi-agent tasks, which include matrix game, predator and prey, and StarCraft Multi-Agent Challenge (Samvelyan et al. 2019). The experimental results demonstrate that our proposed algorithm, OVI, successfully overcomes the RO problem and achieves significant coordination among agents to solve these tasks. In contrast, the baselines struggle in the sub-optimal policies and show poor performance.

Related Work

In this section, we present a review of value-based MARL algorithms that aim to address the RO problem. We divide them into two major categories: independent learners, which focus on decentralized value estimations of agents' local actions, and value-decomposition learners, which concentrate on approximating the true joint action value function.

The canonical independent learner used to solve cooperative multi-agent tasks is decentralized Q-learning (Tan 1993). This algorithm learns average-based value functions for decentralized agents and thus is easy to be biased by other agents' non-cooperative actions, which leads to the RO problem. To overcome this limitation, distributed Q-learning (Lauer and Riedmiller 2000) employs a maximum-based (optimistic) decentralized value function, where each agent assumes that other agents always select their coordinated actions. This encourages agents to identify and execute their local coordinated actions, thus alleviating the RO problem. However, being highly optimistic makes distributed Q-learning vulnerable to stochasticity. Hysteretic Q-learning (Matignon, Laurent, and Le Fort-Piat 2007) avoids high optimism by utilizing two different learning rates to update value functions with positive and negative temporal-difference errors respectively. And lenient learner (Panait, Sullivan, and Luke 2006; Wei and Luke 2016) gradually shifts from maximum-based value estimations to average-based value estimations by employing decreasing lenience. Although these independent learner algorithms effectively coordinate decentralized agents using maximum-based and average-based value estimations in fully observable markov games (where agents have access to the states), achieving satisfactory performance in practice is a non-trivial task due to the limited information available to agents in partially observable tasks (Palmer, Savani, and Tuyls 2018).

In contrast to independent learners, value-decomposition learners employ a centralized controller to guide the policy learning of multiple decentralized agents. Specifically, value-decomposition algorithms learn a factored global action value function that is comprised by all agents' decentralized utility functions. Furthermore, this factorization must adhere to the IGM principle (Son et al. 2019), which enables a tractable greedy search over the factored global action value function through local greedy action selections over the decentralized utility functions. If the factored global action value function accurately approximates the true joint action values, the optimal joint actions can be identified and thus the RO problem is solved. However, the representational capacity of the factored global action value function is limited by the monotonic constraint present in VDN (Sunehag et al. 2017) and QMIX (Rashid et al. 2020b), which results in an ongoing challenge in addressing the RO problem. This limitation can be addressed by works falling into two categories: WQMIX (Rashid et al. 2020a), which places more weights on the optimal joint actions to recover correct value functions for them, and QTRAN (Son et al. 2019) and QPLEX (Wang et al. 2020a), which introduce additional terms to correct the discrepancy between the learned factored global action value functions and the true joint ones, thereby achieving complete expressiveness of IGM-class value functions. These algorithms heavily depend on precise identification of optimal joint actions during the learning process; however, practical outcomes often reveal failures in this regard, leading to poor performance (Gupta et al. 2021). In such scenario, the RO problem persists due to the inaccurate approximation of true joint action value functions.

In this work, we introduce multiple optimistic instructors

to identify and suggest potentially optimal joint actions for value-decomposition learners. These suggestions are then utilized to rectify the learned factored global action value function, which helps to overcome the RO challenge. Thus, our work complements existing value-decomposition algorithms by enhancing their performance through the accurate identification of potentially optimal joint actions and the rectification of the global action value function to recover them.

Preliminary

In this section, we first formalize the tasks in our work. Then, we give an introduction to optimistic value estimation and several value-decomposition algorithms.

Dec-POMDP. A fully cooperative multi-agent task where agents make decisions in a decentralized manner is usually modeled as a Decentralized Partially Observable Markov Decision Process (Dec-POMDP). Dec-POMDP is represented by a tuple $< \mathcal{N}, S, \boldsymbol{A}, \mathcal{R}, \mathcal{P}, \boldsymbol{Z}, \boldsymbol{O}, \gamma >$, where $\mathcal{N} = \{1, 2, \ldots, n\}$ denotes the agent set and S is the state space. $\boldsymbol{A} = \{A^1 \times A^2 \ldots \times A^n\}$ represents the joint action space of all agents, and A^i is the local action space of agent $i \in \mathcal{N}$. At each time step t, each agent i receives its local observation $o_t^i \in Z^i \in \boldsymbol{Z}$ according to its observation function $O^i(o_t^i|s_t) \in \boldsymbol{O}$ and selects its local action $a_t^i \in A^i$. After all agents select actions, the environment transits to the next state s_{t+1} according to the state transition function $\mathcal{P}(s_{t+1}|s_t, \boldsymbol{a_t})$ and provides all agents with the same reward r_t according to the reward function $\mathcal{R}(s_t, \boldsymbol{a_t})$, where $\boldsymbol{a_t} = (a_t^1, a_t^2, \ldots, a_t^n)$ denotes the joint action of all agents. And γ is a discount factor. To overcome the partially observability challenge, each agent i conditions its decentralized policy $\pi^i(a_t^i|\tau_t^i)$ on the local action-observation history $\tau_t^i = (o_o^i, a_0^i, o_1^i, a_1^i, \ldots, o_t^i)$. The goal of all agents is to learn the optimal joint policy $\boldsymbol{\pi^*} = \{\pi^{1,*}, \pi^{2,*}, \ldots, \pi^{n,*}\}$ that maximizes the cumulative discounted rewards $\mathbb{E}_{\boldsymbol{\pi},\mathcal{P}}[\sum_{t=0}^{\infty} \gamma^t r_t]$.

Optimistic Value Estimation. Each agent i is provided with a decentralized instructive value function $Q_I^i(s, a^i)$, which can be regarded as a projection of the true joint action value function $Q^{jt}(s, a^i, a^{-i})$, where a^{-i} represents the joint action of other agents $-i$ except agent i. Specifically, the average-based projection is defined as follows:

$$Q_I^{i,\boldsymbol{\pi}}(s, a^i) = \sum_{a^{-i}} \pi^{-i}(a^{-i}|s) Q^{jt,\boldsymbol{\pi}}(s, a^i, a^{-i}).$$

Here, π^{-i} represents the joint policy of other agents $-i$, and $Q^{jt,\boldsymbol{\pi}}$ denotes the joint value function of a given joint policy $\boldsymbol{\pi} = \{\pi^i, \pi^{-i}\}$. It is obvious that the average-based projection is easily affected by other agents' non-cooperative actions and suffers from the RO problem. In contrast, the optimistic (maximum-based) projection is defined as follows:

$$Q_I^{i,opt}(s, a^i) = \max_{a^{-i}} Q^{jt,*}(s, a^i, a^{-i}),$$

where $Q^{jt,*}$ denotes the joint value function of an optimal joint policy $\boldsymbol{\pi}^*$. For each agent i, this projection implicitly assumes that other agents $-i$ always select their coordinated actions and thus eliminates the negative effects caused by other agents' non-cooperation. As a result, the optimistic instructive value functions can encourage agents to select their local coordinated actions, leading to the optimal joint policy.

Value Decomposition Algorithms. Value-decomposition algorithms aim to learn a factored global action value function, denoted as Q^{total}, that can be factored into all agents' decentralized utility function Q^i. The IGM principle is introduced to ensure the optimal consistency between these two value functions, which is defined as follows:

$$\arg\max_{\boldsymbol{a}} Q^{total}(\boldsymbol{\tau}, \boldsymbol{a}) = \begin{pmatrix} \arg\max_{a^1} Q^1(\tau^1, a^1) \\ \arg\max_{a^2} Q^2(\tau^2, a^2) \\ \ldots \\ \arg\max_{a^n} Q^n(\tau^n, a^n) \end{pmatrix},$$

where $\boldsymbol{\tau}$ denotes the joint action-observation histories of all agents and $\boldsymbol{a}$ is the joint action. Specifically, QMIX adheres to this principle by applying the monotonicity constraint:

$$\frac{\partial Q^{total}(\boldsymbol{\tau}, \boldsymbol{a}, s)}{\partial Q^i(\tau^i, a^i)} \geq 0, \forall i \in \mathcal{N},$$

and achieves it by introducing a mixing network g to calculate the global value function: $Q^{total} = g(Q^1, Q^2, \ldots, Q^n)$. The parameters of g are generated by hyper-networks conditioned on the states, and the weights of g are restricted to be positive to satisfy the monotonicity constraint.

However, the monotonicity constraint is only a sufficient condition for the IGM principle, thus limiting the representational capacity of the factored global action value function. To address this limitation, QPLEX proposes the Advantage-based IGM principle. This principle identifies the necessary condition of the IGM principle and achieves complete expressiveness of the IGM-class factored global action value function by utilizing a duplex dueling architecture.

Methodology

In this section, we provide a comprehensive introduction to our proposed algorithm, *Optimistic Value Instructors (OVI)*. We begin by presenting the motivation behind our algorithm and then delve into the process of learning the optimistic instructive value function for each agent. Additionally, we introduce two instructive constraints aimed at rectifying the learned factored global action value function within the value-decomposition learner. Finally, we summarize the overall learning procedure of our algorithm.

Motivation

Our algorithm is built on the paradigm of value decomposition, which learns multiple utility functions $Q^i(\tau^i, a^i)$ for decentralized agents and composes a factored global action value function $Q^{total}(\boldsymbol{\tau}, \boldsymbol{a})$. Each agent i's utility function is conditioned on its local action-observation history τ^i and local action a^i, which allows for decentralized action selection. The factored global action value function is accessible to the global information (*i.e.*, all agents' joint action-observation histories $\boldsymbol{\tau}$ and joint actions $\boldsymbol{a}$), which enables efficient coordination among agents.

Moreover, the IGM principle is employed to ensure the consistency of optimal action selections between Q^{total} and Q^i, which also enables a tractable search for optimal joint actions. We use f to represent all aggregation functions that adhere to the IGM principle and define Q^{total} as follows:

$$Q^{total}(\boldsymbol{\tau}, \boldsymbol{a}) = f(Q^1(\tau^1, a^1), \ldots, Q^n(\tau^n, a^n)).$$

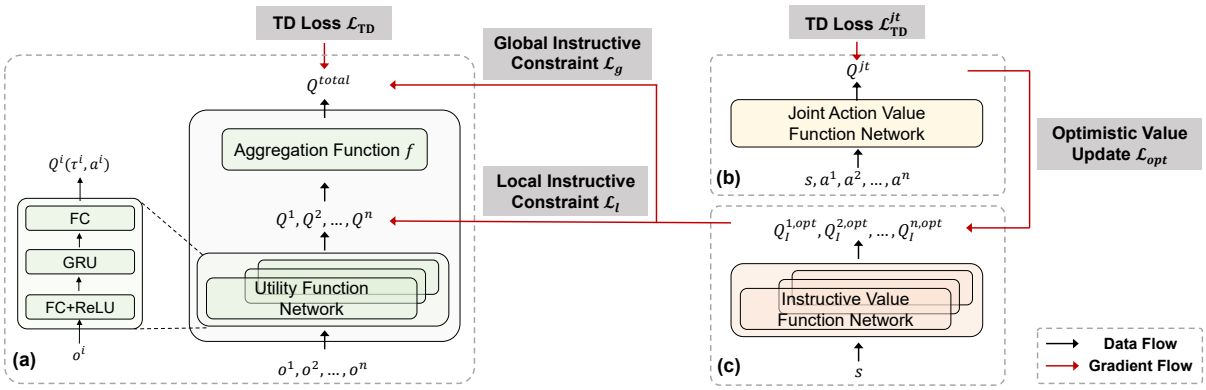

Figure 2: The architecture of OVI. (a) The value-decomposition learner. (b) The joint action value function. (c) The instructors. The value decomposition learner can be instantiated by any multi-agent value-decomposition algorithms.

As mentioned earlier, the representational limitations or inaccurate sampling of the optimal joint actions during the learning process may hinder the learned factored global action value function from accurately approximating the true ones, leading to the RO problem. One promising approach is to assign greater importance to superior joint actions (*i.e.,* accurately identify the optimal joint actions and approximate their value functions). However, directly searching for the optimal joint actions in the joint action space is intractable, and inaccurate sampling can result in sample inefficiency.

To overcome this limitation, we introduce multiple instructors that utilize the optimistic value estimations to customize instructive value functions for agents. The per-agent instructive value function estimates the local action values for each agent based on the optimistic assumption that other agents will always cooperate. Consequently, this function can be used to identify and suggest local cooperative actions that have optimal values for each agent. Based on the suggested optimal joint actions, we rectify the factored global action value function to approximate the true values of these optimal joint actions, thus accurately recovering them.

Optimistic Instructors

For each agent i, we aim to introduce an instructor capable of disregarding other agents' non-cooperation and accurately identifying the agent's local optimal action. We achieve this by learning an optimistic instructive value function $Q_I^{i,opt}(s, a^i)$, which is conditioned on the state s and the agent's local action a^i. The optimistic instructive value function can be updated using a special Q-learning rule, which updates the value $Q_I^{i,opt}(s, a^i)$ only when an increase occurs. Specifically, we define this update as follows:

$$Q_I^{i,opt}(s_t, a_t^i) \leftarrow \begin{cases} Q_I^{i,opt}(s_t, a_t^i) + \delta_t^i & \text{if } \delta_t^i \geq 0 \\ Q_I^{i,opt}(s_t, a_t^i) & \text{else} \end{cases}, \quad (1)$$

where δ_t^i denotes the temporal difference error (TD-error). It is defined as:

$$\delta_t^i = \underbrace{r_t + \gamma \max_{\boldsymbol{a_{t+1}}} Q^{jt}(s_{t+1}, \boldsymbol{a_{t+1}})}_{\text{learning target calculated by } Q^{jt}} - Q_I^{i,opt}(s_t, a_t^i). \quad (2)$$

Here, $Q^{jt}(s, \boldsymbol{a})$ denotes an extra joint action value function conditioned on the state s and all agents' joint action $\boldsymbol{a} = (a^i, a^{-i})$. And we update it with standard Q-learning rule:

$$Q^{jt}(s_t, \boldsymbol{a_t}) \leftarrow (1-\alpha) Q^{jt}(s_t, \boldsymbol{a_t}) + \alpha(r_t + \gamma \max_{\boldsymbol{a_{t+1}}} Q^{jt}(s_{t+1}, \boldsymbol{a_{t+1}})), \quad (3)$$

where α denotes the learning rate.

In the optimistic update above, we separately learn a joint action value function and utilize it to calculate the learning target of the instructive value function. This setting is different from classic Distributed Q-learning, which solely relies on each agent's decentralized value function for the entire optimistic update process. In scenarios where multiple agents strongly influence each other, a fully decentralized value update may lead to instability and poor convergence.

In contrast, we decompose the optimistic update into two distinct sub-processes. The first is that we learn a separate joint action value function that has access to the global information, which is capable of shaping stable learning targets for the instructive value functions of multiple agents. The second is that each agent's instructive value function acts like a supervised learner throughout the learning process, where it is disentangled from the traditional bootstrapping learning target calculation and strives to regress the target determined by the stable joint action value function. By implementing these two sub-processes, we ensure the stable updating of all agents' instructive value function.

Practical Implementations. However, there are still two issues in above process. One is the intractable greedy action search with respect to the joint action value function Q^{jt}. In Equations (2) and (3), the max operation over the large joint action space is intractable in practice. To overcome this limitation, we use the factored global action value function Q^{total} learned by vanilla value-decomposition learners to approximately sample the optimal joint action at next time step, which can be tractably achieved by local greedy searches over the utility functions of individual agents.

The other is the excessive utilization of samples with positive TD-errors. When we use a neural network as the instructive value function approximation, samples with pos-

itive TD-errors may result from the neural network rather than the true reward. Before the neural network is well fitted, such inaccurate samples may be generated frequently and interfere with each agent's instructive value function learning.

Therefore, we also use samples with negative TD-errors to complement the optimistic instructive value functions. Specifically, we employ a complement factor $\beta<1$ to control the degree of optimism. Thus, we redefine the optimistic update rule as follows:

$$Q_I^{i,opt}(s_t, a_t^i) \leftarrow \begin{cases} Q_I^{i,opt}(s_t, a_t^i) + \delta_t^i & \text{if } \delta_t^i \geq 0 \\ Q_I^{i,opt}(s_t, a_t^i) + \beta\delta_t^i & \text{else} \end{cases}, \quad (4)$$

where δ_t^i can be calculated as follows:

$$\begin{aligned} \delta_t^i &= r_t + \gamma Q^{jt}(s_{t+1}, \hat{\boldsymbol{a}}_{\boldsymbol{t+1}}) - Q_I^{i,opt}(s_t, a_t^i), \\ \hat{\boldsymbol{a}}_{\boldsymbol{t+1}} &= \text{argmax}_{\boldsymbol{a_{t+1}}} Q^{total}(\boldsymbol{\tau_{t+1}}, \boldsymbol{a_{t+1}}). \end{aligned} \quad (5)$$

Similarly, the joint action value function Q^{jt} is updated as follows:

$$\begin{aligned} Q^{jt}(s_t, \boldsymbol{a_t}) \leftarrow &(1-\alpha)Q^{jt}(s_t, \boldsymbol{a_t}) + \\ &\alpha(r_t + \gamma Q^{jt}(s_{t+1}, \hat{\boldsymbol{a}}_{\boldsymbol{t+1}})), \\ \hat{\boldsymbol{a}}_{\boldsymbol{t+1}} &= \text{argmax}_{\boldsymbol{a_{t+1}}} Q^{total}(\boldsymbol{\tau_{t+1}}, \boldsymbol{a_{t+1}}). \end{aligned} \quad (6)$$

Instructive Constraints

Based on the learned optimistic instructive value function, we present an instructive policy for each agent i, which is defined as follows:

$$\pi_I^i(a^i|s) = \frac{\exp(Q_I^{i,opt}(s, a^i))}{\sum_{a^i} \exp(Q_I^{i,opt}(s, a^i))}, \forall a^i \in A^i. \quad (7)$$

Note that although the instructive value function of each agent and the corresponding instructive policy promise to identify the optimal local actions for the agent, directly accessing the state s violates the decentralized execution when agents are located in partially observable multi-agent tasks. In addition, the value-decomposition learner often suffers from representational limitations or inaccurate sampling of optimal joint actions during the learning process, preventing it from accurately approximating the true joint action value function and leading to the RO problem. Therefore, we intend to use the instructive policies of all agents to assist the value-decomposition learner during centralized learning. For this purpose, we propose two instructive constraints to rectify the factored global action value function in the value-decomposition learner, aiding in the accurate recovery of the optimal joint actions.

Global Instructive Constraint. Specifically, for value-decomposition learners that inherently suffer from representational limitations, we introduce the global instructive constraint to guide the global value function learning as follows:

$$\begin{aligned} \mathcal{L}_g &= \min \sum_{k=1}^{b} \omega(s_t, \boldsymbol{a_t})(Q^{total}(\boldsymbol{\tau_t}, \boldsymbol{a_t}) - y_k)^2, \\ \omega(s_t, \boldsymbol{a_t}) &= \begin{cases} 1 & \text{if } \boldsymbol{a_t} = \hat{\boldsymbol{a}}_{\boldsymbol{t}} \text{ or } y_k > Q^{jt}(s_t, \hat{\boldsymbol{a}}_{\boldsymbol{t}}) \\ \lambda & \text{else} \end{cases}, \end{aligned} \quad (8)$$

where $y_k = r_t + \gamma Q^{jt}(s_{t+1}, \text{argmax}_{\boldsymbol{a_{t+1}}} Q^{total}(\boldsymbol{\tau_{t+1}}, \boldsymbol{a_{t+1}}))$ and b denotes batches sampled from the replay buffer. λ is the correcting weight. In addition, $\hat{\boldsymbol{a}}_{\boldsymbol{t}} = (\hat{a}_t^1, \hat{a}_t^2, \ldots, \hat{a}_t^n)$ denotes the suggested optimal joint actions by all agents' instructive policies, where $\hat{a}_t^i = \text{greedy}_{a^i \in A^i} \pi_I^i(a^i|s_t)$.

Intuitively, the global instructive constraint encourages the learned factored global action value function to place more weights on samples with suggested optimal joint actions or potentially better joint actions (indicated by $y_k > Q^{jt}(s_t, \hat{\boldsymbol{a}}_{\boldsymbol{t}})$). With this biased value function learning, the global action value function can precisely approximate the values of these optimal joint actions, enabling agents to recover them using decentralized utility functions. This constraint shares similarities with Centrally-Weighted QMIX (CW-QMIX). However, the key difference is that the optimistic instructors accurately suggest optimal joint actions rather than approximate sampling, thereby accelerating the learning of optimal joint policies.

Local Instructive Constraint. Although some recent value-decomposition learners address the issue of representation limitations by implementing necessary conditions of the IGM principle, they usually perform poorly due to the inaccurate sampling of optimal joint actions during the learning process. To assist them, we propose the local instructive constraint that guides each agent with their respective instructive policies:

$$\mathcal{L}_l = \min D_{\text{KL}}(\pi^i(\cdot|\tau^i)||\pi_I^i(\cdot|s)) \quad \forall i \in \mathcal{N}, \quad (9)$$

where $\pi^i(\cdot|\tau^i)$ is a Boltzmann policy with respect to each agent's decentralized utility function defined as:

$$\pi^i(a^i|\tau^i) = \frac{\exp(Q^i(\tau^i, a^i))}{\sum_{a^i} \exp(Q^i(\tau^i, a^i))}, \forall a^i \in A^i. \quad (10)$$

The intuition behind this constraint is that each agent's instructive policy is capable of identifying optimal local action and we make each agent's decentralized policy with respect to the utility function imitate it, which encourages agents to select their optimal local actions. As a result, the value-decomposition learner can accurately approximate the values of optimal joint actions based on the full expressiveness of the IGM principle and enable a tractable search for the optimal joint actions through agents' decentralized selections.

Overall Learning Procedure

As shown in Figure 2, our algorithm consists of four components, an aggregation function f that generates the factored global action-value function Q^{total}, each agent's utility function Q^i, the joint action-value function Q^{jt}, and each agent's instructive value function $Q_I^{i,opt}$. All modules are implemented by neural networks and shared among agents to facilitate efficient policy learning. For value decomposition learners that suffer from representational limitations, we define the learning objective of OVI as follows:

$$\mathcal{L} = \mathcal{L}_{\text{TD}}^{jt} + \mathcal{L}_g + \mathcal{L}_{opt}. \quad (11)$$

For value-decomposition learners that achieve full expressiveness of the IGM principle, we employ the learning objective below:

$$\mathcal{L} = \mathcal{L}_{\text{TD}} + \mathcal{L}_{\text{TD}}^{jt} + \mathcal{L}_l + \mathcal{L}_{opt}. \quad (12)$$

A^1 \ A^2	a^1	a^2	a^3
a^1	**8**	-12	-12
a^2	-12	6	0
a^3	-12	0	6

(a) Payoff matrix

A^1 \ A^2	$a^1(8.00)$	$a^2(1.99)$	$a^3(2.30)$
$a^1(8.00)$	**8.12**	-11	-11
$a^2(1.93)$	-11	-11	-11
$a^3(2.28)$	-11	-11	-11

(b) Payoff learned by QMIX-OVI

A^1 \ A^2	$a^1(7.99)$	$a^2(2.86)$	$a^3(0.41)$
$a^1(7.99)$	**7.96**	-13	-18
$a^2(2.87)$	4.24	-17	-21
$a^3(0.44)$	3.49	-18	-22

(c) Payoff learned by QPLEX-OVI

Table 1: The payoff matrix and the value functions (payoffs) learned by QMIX-OVI and QPLEX-OVI.

In Equations (11) and (12), $\mathcal{L}_{\text{TD}}$ and $\mathcal{L}_{\text{TD}}^{jt}$ respectively denote the update losses of the factored global action value function and the separately learned joint ones, given by:

$$\mathcal{L}_{\text{TD}} = \sum_{k=1}^{b} (Q^{total}(\boldsymbol{\tau}_t, \boldsymbol{a_t}) - y_k)^2,$$

$$\mathcal{L}_{\text{TD}}^{jt} = \sum_{k=1}^{b} (Q^{jt}(s_t, \boldsymbol{a_t}) - y_k)^2,$$

where $y_k = r_t + \gamma Q^{jt}(s_{t+1}, \text{argmax}_{\boldsymbol{a_{t+1}}} Q^{total}(\boldsymbol{\tau_{t+1}}, \boldsymbol{a_{t+1}}))$ is the learning target and b are sampled batches. $\mathcal{L}_g$ and $\mathcal{L}_l$ denote the global and local instructive constraints, defined by Equations (8) and (9). $\mathcal{L}_{opt}$ represents the loss of each agent's optimistic instructive value function defined by:

$$\mathcal{L}_{opt} = \sum_{k=1}^{b} \sum_{i \in \mathcal{N}} w^i {\delta^i}^2, \; w^i = 1 \text{ if } \delta^i \geq 0 \text{ else } \beta,$$

where δ^i is defined by Equation (5). In addition, we utilize target networks to enable stability. More details of our algorithm can be found in Appendix A.

Experiment

In this section, we incorporate OVI into QMIX and QPLEX, and compare our algorithms (QMIX-OVI and QPLEX-OVI) with state-of-the-art baselines: CW-QMIX (Rashid et al. 2020a), OW-QMIX (Rashid et al. 2020a), QPLEX (Wang et al. 2020a), QTRAN (Son et al. 2019) and MAVEN (Mahajan et al. 2019). For fair evaluation, all experimental results are illustrated with the median performance and the standard error over five random seeds. More details about algorithmic implementations and experimental settings are provided in Appendix B.

Matrix Game

We begin by evaluating all algorithms on a matrix game shown in Table 1a, where two agents must perform the optimal joint action (a^1, a^1) to receive the best reward. However, the sub-optimal joint actions (a^2, a^2) and (a^3, a^3) make decentralized agents prefer to select their local actions a^2 or a^3 when other agents select actions uniformly during learning, which leads to the RO problem.

Figure 3 shows the comparison with baselines in the matrix game. We can observe that QMIX fails to select the optimal joint action and struggles in sub-optimal ones, leading to a reward of 6. QPLEX and MAVEN occasionally make it and thus suffer from large variance of received rewards.

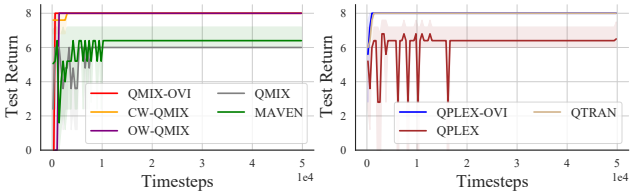

Figure 3: Comparison results in the matrix game.

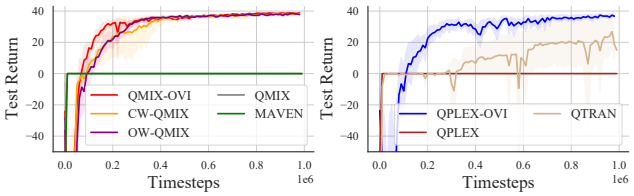

Figure 4: Comparison results in the predator and prey.

In contrast, our algorithms (QMIX-OVI, QPLEX-OVI) succeed in suggesting and selecting the optimal joint action, which demonstrates superior performance than other baselines (CW-QMIX, OW-QMIX, QTRAN).

Also we analyze the value functions learned by our algorithms. As shown in Table 1b and 1c, each agent's instructive value estimations of its local actions (*i.e.,* 8.00 for a^1 in QMIX-OVI and 7.99 for a^1 in QPLEX-OVI) adhere to the optimistic property, which equal to the highest cooperative rewards. Based on these optimistic instructors, the factored global action-value function is rectified to accurately approximate the optimal joint actions' values (*i.e.,* 8.12 for joint action (a^1, a^1) in QMIX-OVI and 7.96 in QPLEX-OVI), which leads to the optimal joint action selection.

Predator and Prey

To further evaluate the effectiveness of OVI in addressing the RO problem, we conduct experiments in the predator and prey, a partially-observable task involving 8 predators and 8 preys. A negative reward -2 is emitted when a single predator tries to capture a prey alone and $+10$ is received when two predators cooperate to capture a prey.

As shown in Figure 4, QMIX, QPLEX and MAVEN all fail to learn effective policies to capture the preys. Although QTRAN benefits from the complete IGM expressiveness, its poor performance and high variance suggest difficulties in approximating the true joint action value function. Compared with CW-QMIX and OW-QMIX, the optimistic in-

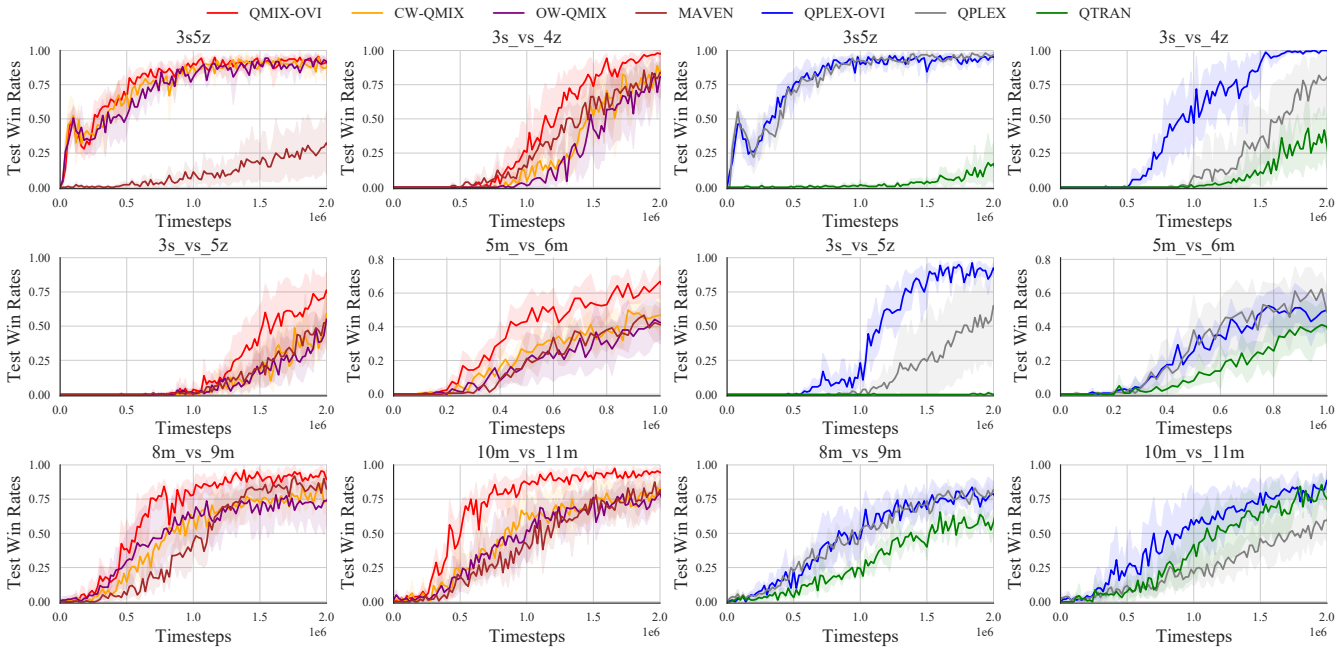

Figure 5: Comparison results on six maps in the StarCraft Multi-Agent Challenge.

structors in OVI accurately suggest the optimal joint actions and thus QMIX-OVI solves this tasks faster. In addition, based on OVI, predators learned by QPLEX-OVI succeed in learning cooperative policies that achieve best returns, which further demonstrates the effectiveness of OVI.

StarCraft Multi-Agent Challenge

To further demonstrate the scalability of OVI in more complex domains, we evaluate it using the StarCraft Multi-Agent Challenge (SMAC) benchmark. Our evaluation focuses on six maps, namely 3s5z, 3s_vs_4z, 3s_vs_5z, 5m_vs_6m, 8m_vs_9m, and 10m_vs_11m, which serve as our testbeds.

Figure 5 shows the performance of all algorithms on these selected maps. One can observe that our algorithm QMIX-OVI significant ourperforms the counterpart baselines (CW-QMIX and OW-QMIX) and quickly learns cooperative policy to achieve high testing win rates on all maps. CW-QMIX and OW-QMIX also employ the optimistic belief to identify potentially optimal joint actions. However, they directly search on the large joint action space by validating whether the TD-errors of sampled joint actions are positive or not. Compared to our proposed optimistic instructors, the inaccurate search over the joint action space may lead to sample inefficiency. As a result, CW-QMIX and OW-QMIX suffer from slow learning speed and achieve low testing win rates. QPLEX-OVI achieves significant improvement in convergence performance on 3s_vs_4z, 3s_vs_5z and 10m_vs_11m compared to its backbone QPLEX, which may be attributed to the frequent optimal joint action selections caused by local instructive constraint. However, QPLEX-OVI achieves similar performance to QPLEX on other maps. We hypothesis that QPLEX is capable of identifying the optimal policies due to its complete IGM expressiveness achieved by the duplex dueling architecture on these maps, and thus the benefit brought by OVI is not obvious. In addition, QTRAN and MAVEN perform worst and fail to defeat enemies.

Ablation Studies

To examine the impact of factor β on OVI's performance, we set β to 0.01, 0.1, 0.3, 0.5 and 0.7 for QMIX-OVI and QPLEX-OVI, which are used as multiple baselines. The experimental results, as presented in Appendix C, demonstrate that OVI with small β achieves superior performance. As β approaches 1, the optimistic instructive value update decreases, transitioning to average-based updates. This change leads to the RO problem and thus degrades the performance.

Conclusion

Decentralized agents learned by value-decomposition algorithms suffer from the RO problem due to the representational limitations or the inaccurate sampling of optimal joint actions during learning. This paper introduces multiple optimistic instructors to accurately suggest optimal joint actions and rectifies the factored global action value function to recover these optimal actions, thus addressing this problem.

Limitations and Future Work. When agents' joint action space is extremely large in complex multi-agent tasks, separately learning the joint action value function may lead to sample inefficiency. This issue may be alleviated by employing the game abstraction technique to simplify the task or introducing prior knowledge to accelerate each agent's policy learning. In addition, we also aim to incorporate the optimistic instructors into policy gradient paradigm to help improve the performance. We leave them as our future work.

Acknowledgments

This work is supported in part by Science and Technology Innovation 2030 New Generation Artificial Intelligence Major Project (No.2018AAA0100905), the National Natural Science Foundation of China (No.62192783, No.62106100, No.62276142), Primary Research & Development Plan of Jiangsu Province (No.BE2021028), Jiangsu Natural Science Foundation (BK20221441), Jiangsu Provincial Double-Innovation Doctor Program (JSSCBS20210021), Young Elite Scientists Sponsorship Program by CAST (2023QNRC001) and in part by the Collaborative Innovation Center of Novel Software Technology and Industrialization. What's more, the authors greatly thank all anonymous reviewers for their valuable comments to this work.

References

Cao, Y.; Yu, W.; Ren, W.; and Chen, G. 2012. An overview of recent progress in the study of distributed multi-agent coordination. *IEEE Transactions on Industrial informatics*, 9(1): 427–438.

Claus, C.; and Boutilier, C. 1998. The dynamics of reinforcement learning in cooperative multiagent systems. *AAAI/IAAI*, 1998(746-752): 2.

Gupta, T.; Mahajan, A.; Peng, B.; Böhmer, W.; and Whiteson, S. 2021. Uneven: Universal value exploration for multi-agent reinforcement learning. In *International Conference on Machine Learning*, 3930–3941. PMLR.

Huang, Y.; Wu, S.; Mu, Z.; Long, X.; Chu, S.; and Zhao, G. 2020. A multi-agent reinforcement learning method for swarm robots in space collaborative exploration. In *2020 6th international conference on control, automation and robotics (ICCAR)*, 139–144. IEEE.

Lauer, M.; and Riedmiller, M. A. 2000. An Algorithm for Distributed Reinforcement Learning in Cooperative Multi-Agent Systems. In *Proceedings of the Seventeenth International Conference on Machine Learning*, 535–542.

Mahajan, A.; Rashid, T.; Samvelyan, M.; and Whiteson, S. 2019. Maven: Multi-agent variational exploration. *Advances in neural information processing systems*, 32.

Matignon, L.; Laurent, G. J.; and Le Fort-Piat, N. 2007. Hysteretic q-learning: an algorithm for decentralized reinforcement learning in cooperative multi-agent teams. In *2007 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 64–69. IEEE.

Palmer, G.; Savani, R.; and Tuyls, K. 2018. Negative update intervals in deep multi-agent reinforcement learning. *arXiv preprint arXiv:1809.05096*.

Panait, L.; Sullivan, K.; and Luke, S. 2006. Lenient learners in cooperative multiagent systems. In *Proceedings of the fifth international joint conference on Autonomous agents and multiagent systems*, 801–803.

Rashid, T.; Farquhar, G.; Peng, B.; and Whiteson, S. 2020a. Weighted qmix: Expanding monotonic value function factorisation for deep multi-agent reinforcement learning. *Advances in neural information processing systems*, 33: 10199–10210.

Rashid, T.; Samvelyan, M.; De Witt, C. S.; Farquhar, G.; Foerster, J.; and Whiteson, S. 2020b. Monotonic value function factorisation for deep multi-agent reinforcement learning. *The Journal of Machine Learning Research*, 21(1): 7234–7284.

Samvelyan, M.; Rashid, T.; De Witt, C. S.; Farquhar, G.; Nardelli, N.; Rudner, T. G.; Hung, C.-M.; Torr, P. H.; Foerster, J.; and Whiteson, S. 2019. The starcraft multi-agent challenge. *arXiv preprint arXiv:1902.04043*.

Son, K.; Kim, D.; Kang, W. J.; Hostallero, D. E.; and Yi, Y. 2019. Qtran: Learning to factorize with transformation for cooperative multi-agent reinforcement learning. In *International conference on machine learning*, 5887–5896. PMLR.

Sunehag, P.; Lever, G.; Gruslys, A.; Czarnecki, W. M.; Zambaldi, V.; Jaderberg, M.; Lanctot, M.; Sonnerat, N.; Leibo, J. Z.; Tuyls, K.; et al. 2017. Value-decomposition networks for cooperative multi-agent learning. *arXiv preprint arXiv:1706.05296*.

Tan, M. 1993. Multi-agent reinforcement learning: Independent vs. cooperative agents. In *Proceedings of the tenth international conference on machine learning*, 330–337.

Wang, J.; Ren, Z.; Liu, T.; Yu, Y.; and Zhang, C. 2020a. Qplex: Duplex dueling multi-agent q-learning. *arXiv preprint arXiv:2008.01062*.

Wang, X.; Ke, L.; Qiao, Z.; and Chai, X. 2020b. Large-scale traffic signal control using a novel multiagent reinforcement learning. *IEEE transactions on cybernetics*, 51(1): 174–187.

Wei, E.; and Luke, S. 2016. Lenient learning in independent-learner stochastic cooperative games. *The Journal of Machine Learning Research*, 17(1): 2914–2955.

Zhang, C.; and Lesser, V. 2011. Coordinated multi-agent reinforcement learning in networked distributed POMDPs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 25, 764–770.